DOCUMENT RESUME

ED 189 139 TM 800 309

AUTHOR Ree, Malcom James; Jensen, Harald E.
TITLE Item Characteristic Curve Parameters: Effects of Sample Size on Linear Equating.
INSTITUTION Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO AFHRL-TR-79-70
PUB DATE Feb 80
NOTE 16p.

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Computer Assisted Testing; *Equated Scores; Error of Measurement; Item Analysis; Item Banks; *Latent Trait Theory; *Reliability; *Sampling; Simulation; *Test Items
IDENTIFIERS Anchor Items Method

ABSTRACT

By means of computer simulation of test responses, the reliability of item analysis data and the accuracy of equating were examined for hypothetical samples of 250, 500, 1000, and 2000 subjects for two tests with 20 equating items plus 60 additional items on the same scale. Birnbaum's three-parameter logistic model was used for the simulation. The equating procedure used was described as the Anchor Items Method, and the equations for this method are given. The results of the simulation are presented in tables showing the following data for all four sample sizes: correlations between known and estimated item parameters for all three parameters, and sums and means of absolute differences for differences (errors) for all three parameters. Also shown are means, standard deviations, and correlations for the a and b equating parameters. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

AFHRL-TR-79-70

AIR FORCE HUMAN RESOURCES LABORATORY

ITEM CHARACTERISTIC CURVE PARAMETERS: EFFECTS OF SAMPLE SIZE ON LINEAR EQUATING

By

Malcolm James Ree
Harald E. Jensen

PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

February 1980
Interim Report for Period April 1979 – June 1979

Approved for public release; distribution unlimited.

ED189139

TM800309

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This interim report was submitted by Personnel Research Division, under project 7719, with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Dr. Malcolm James Ree was the Principal Investigator for the Laboratory.

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

NANCY GUINN, Technical Director
Personnel Research Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFHRL-TR-79-70

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*: ITEM CHARACTERISTIC CURVE PARAMETERS: EFFECTS OF SAMPLE SIZE ON LINEAR EQUATING

5. TYPE OF REPORT & PERIOD COVERED: Interim, April 1979 – June 1979

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Malcolm James Ree, Harald E. Jensen

8. CONTRACT OR GRANT NUMBER(s)

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Personnel Research Division, Air Force Human Resources Laboratory, Brooks Air Force Base, Texas 78235

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 62703F, 77191522

11. CONTROLLING OFFICE NAME AND ADDRESS: HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235

12. REPORT DATE: February 1980

13. NUMBER OF PAGES: 16

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*

15. SECURITY CLASS. *(of this report)*: Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

item analysis
latent-trait theory
sample size
simulations
adaptive testing
computer adaptive testing
equating
item characteristic curve

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

In order to establish large item pools, it is necessary to try out hundreds of items for possible inclusion. Because it is impractical to administer several hundred items to several thousand people, the invariance properties of Item Characteristic Curve (ICC) parameters are used to advantage to equate test items. A simulation study was conducted in order to investigate the effects of errors in estimated ICC parameters on equated parameters. Four samples of simulated examinees, ranging from 250 to 2,000 subjects, were administered a different test containing the 20 anchor items, plus 60 additional items on the same scale. The effectiveness of the equating procedures was studied through analysis of summed absolute deviations from the known item parameters of the estimated item parameters and the equated item parameters for the various combinations of "anchor" item calibration sample size and equating same size.

DD FORM 1 JAN 73 1473

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

MAR 2 1 1980

PREFACE

This research was conducted under Project 7719, Air Force Development of Selection, Assignment, Performance Evaluation, Retention and Utilization Devices; Task 771915, Perceptual and Computer Managed Measurement.

The authors wish to express their appreciation to Dr. Janos Koplyay, Air Force Human Resources Laboratory and to Mr. Thomas Warm, U.S. Coast Guard Institute, for their critical review of the effort.

TABLE OF CONTENTS

		Page
I.	Introduction	5
	ICC Parameters	5
	Linking Paradigms	7
II.	Method	7
	Generation of Item Responses	8
III.	Results	9
IV.	Discussion	13
	Estimating and Equating a	13
	Estimating and Equating b	13
	Estimating and Equating c	14
References		14

LIST OF ILLUSTRATIONS

Figure		Page
1	Item characteristic curves	6
2	Error in Estimation of ICC Parameter	11

LIST OF TABLES

Table		Page
1	Mean, Standard Deviation, Minimum and Maximum of θ for Groups S1 and S2	8
2	Means and Standard Deviations of the Generated Item Parameters for Test 1 (T1) and Test 2 (T2)	8
3	ICC Parameters of the 20 Anchor Items Common to Both Tests	9
4	Intercorrelations between Known and Estimated ICC Parameters for Both Groups with Varying Sample Sizes	10
5	Summed Absolute Deviations (Σ|Error|) and Averaged Absolute Deviations ($\overline{\|Error\|}$) for the Three ICC Parameters for the Two Tests	10
6	Summed Absolute Deviations (Σ|Error|) and Average Absolute Deviations ($\overline{\|Error\|}$) for the a and be Parameters for Various Equating and Calibrating Sample Sizes	12
7	Intercorrelations, Means, and Standard Deviation of the Estimated a Parameters for Test 2	12
8	Intercorrelations, Means, and Standard Deviation of the Estimated b Parameters for Test 2	13

ITEM CHARACTERISTIC CURVE PARAMETERS: EFFECTS OF SAMPLE SIZE ON LINEAR EQUATING

I. INTRODUCTION

The application of the technology of computer driven adaptive testing requires the development of large banks of test items. Each bank may contain 250 to 400 items, and all must measure the same ability on the same metric or scale. It is unreasonable and impracticable to assemble a single group of 2,000 subjects for 250 to 400 minutes to try all the items; therefore, a method for linking together subsets of items administered to varying groups must be investigated. Item Characteristic Curve (ICC) theory offers a unique method of linking subsets of test items due to the invariance property of the ICC parameters. This invariance property rests on the two major theoretical assumptions of latent-trait theory: (a) unidimensionality and (b) local independence. Unidimensionality means that only a single ability is being measured and is assumed to be the property of an item pool, even when assembled into subsets. Local independence means that the subjects' responses to an item are independent of the responses to another item. More simply put, this means that the item response is a function of ability and no other factor. In effect, this is a restatement of the unidimensionality assumption. If an item pool is unidimensional, then any shift in score metric that is due to a linear transformation may be corrected or adjusted by application of the proper complementary linear transformation. This is what is meant by the idea that latent-trait parameters are invariant to a linear transformation, and it is this theoretical property that allows item pools to be linked and transformed to a common metric. In previous research efforts, item pools have been linked via the method of linear equating (see Lord, 1977; Ree, 1977; Sympson & Ree, in press) with apparent success. To date, there has been little research on the efficacy of these linking procedures and the effects of errors in ICC parameter estimation on their (linearly) transformed values.

ICC Parameters

The three parameter logistic model of Birnbaum (Lord & Novick, 1968) is the most frequently used for relating item responses to subjects' ability. The three parameters, a, b, and c, are item discrimination, item difficulty (or location), and probability of chance success (or lower asymptote), respectively.

The curve described by these parameters takes the shape of an ogive (cumulative frequency) or an "s" with the upper asymptote approaching a probability of 1.0 and usually a lower asymptote of a probability greater than 0.0. The ogive describes the probability of obtaining a correct answer to an item as a monotonic increasing function of ability.

The item discrimination parameter, a, is a function of the slope of the ICC and generally ranges from .5 to about 2.5. The value of a equal to about 1.0 is typical of many test items, while a values below .5 are insufficiently discriminating for most testing purposes, and a values above 2.0 are infrequently found.

The item difficulty parameter, b, describes the point of inflection of the ICC and is usually scaled between -2.5 and $+2.5$, although the metric is arbitrary.

The item guessing parameter, c, is the lower asymptote of the ICC and is generally conceived as the probability of selecting the correct item-option by chance alone. Most test items have c parameters greater than 0.0 and less than or equal to .30.

Figure 1 shows three ICCs. The horizontal axis is scaled in units of ability θ and the vertical axis is the probability of answering the item correctly. The solid curved line shows an ICC

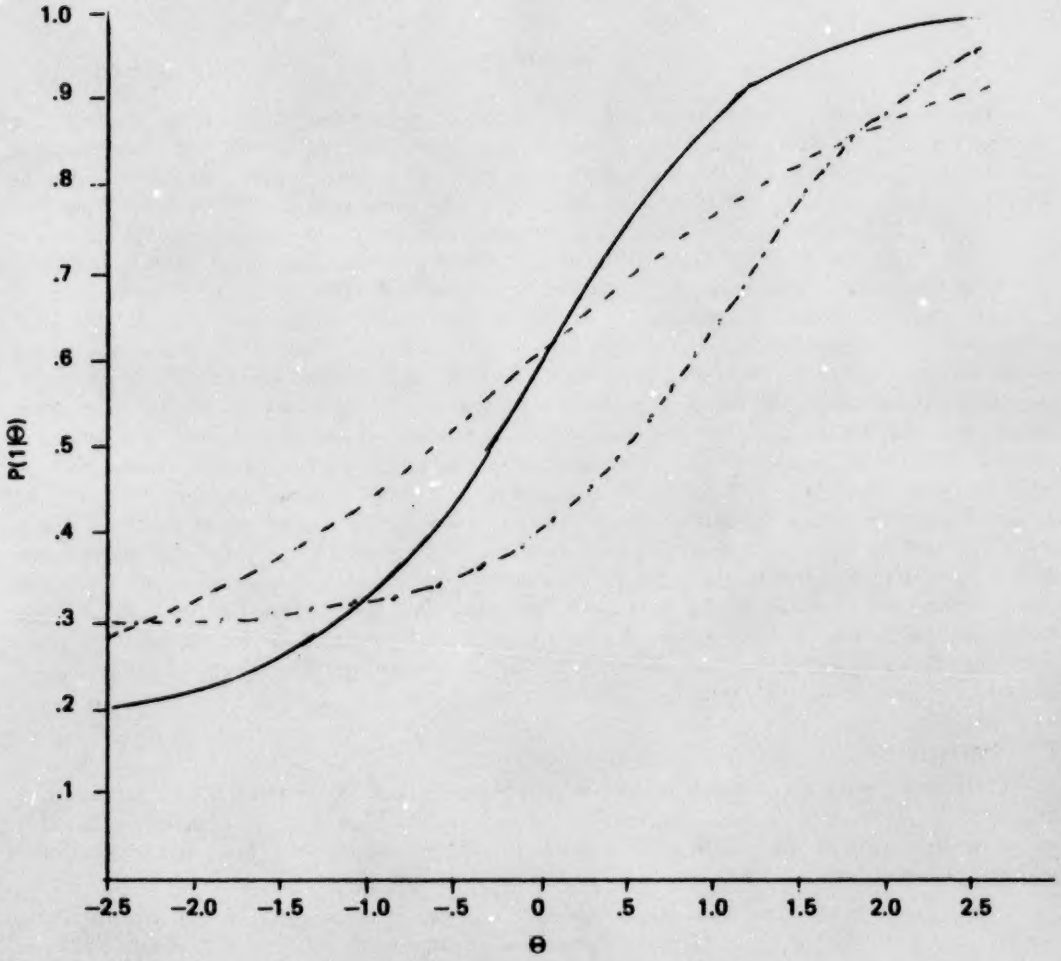

Figure 1. **Item characteristic curves.**

for an item of average difficulty with acceptable discrimination and the lower asymptote appropriate for a five-item multiple-choice item. The dashed line shows an item of identical difficulty, c value of .28, but with a lower a value. Note how the slope of the curve is less steep. The third curve, dot-dash line, shows an item with a c value of .30, an a parameter of 1.0, and the b parameter equal to 1.0. As the b parameter changes, the location of the inflection point of the curve is displaced along the horizontal axis.

Equation 1 presents the mathematical function describing the curve.

$$p(\theta)_j = c_i + (1 - c_i)\,(1 + e^{(-1.7a_i\,(\theta - b_i))})^{-1} \qquad (1)$$

Previous research (Ree, 1978) indicates that the ICC parameters may be estimated with some reasonable degree of accuracy, providing a sufficient sample of examinees with an appropriate distribution of ability, θ is available.

Linking Paradigms

Two fundamental linking procedures may be defined and are known as the Anchor Items Method (AIM) and the Anchor Subjects Method (ASM). In AIM, every subset of items is administered to a different sample of subjects, but embedded into the group of items to be analyzed is a common (or anchor) set of items. During analysis, the anchor items are identified, and the following linear transformation is applied to the resultant ICC parameters:

$$b_t = \left(\frac{sb_t}{sb_2}\right) b_2 + \left[\bar{b}_t - \left(\frac{sb_t}{sb_2}\right) \bar{b}_2\right] \quad (2)$$

Where b_t is the item location parameter transformed to the desired scale and sb_t and sb_2 are standard deviations of the desired scale and observed scale respectively. A similar procedure for the *a* parameter is defined by

$$a_t = a_2 \cdot \frac{sb_2}{sb_t} \quad (3)$$

Where a_t is the item discrimination parameter transformed to the desired scale, a_2 is the observed *a* parameter, and sb_t and sb_2 are as in equation (2). Because the *c* parameter is measured on the probability axis, it does not change and no transformation need be applied.

The ASM requires that the same group of subjects be available to take each subset of items. It is extremely unlikely that the same 2,000 subjects could be assembled to take items over a long period of time as would be required to place tests on the same metric from year to year. For this reason, the ASM method seems less likely to find long-term practical application. Because of its potential for use, the AIM procedure is the subject of the present study.

II. METHOD

In order to have a known standard for reference, a simulation study was run using two groups of subjects, a single set of 20 anchor items and two differing groups of 60 experimental, or non-anchor, items. These two groups of items were assembled into two tests designated T1 and T2. Both groups of simulated subjects were specified to have about the same normal distribution of θ. Table 1 shows the mean, standard deviation, minimum and maximum of θ for the groups S1 and S2. These two groups represent what might be expected if subjects for experimental testing were picked from some larger pool, such as candidates for military enlistment for example. Response vectors for these subjects were generated on the two tests.

Table 1. **Mean, Standard Deviation, Minimum and Maximum of θ for Groups S1 and S2**

Parameter	Groups S1	Groups S2
$\bar{X}_\theta$	−0.0145	0.0250
σ_θ	0.9976	1.0045
Minimum	−2.6000	−2.6000
Maximum	2.6000	2.6000

Generation of Item Responses

In order to generate a vector of item responses for each "subject" the θ values were used in equation (1) to compute the likelihood of "passing" each item.

Because Equation 1 yields a number $P(\theta)_j$ such that $0.0 < P(\theta)_j < 1.0$, a number X_j is drawn from a uniform (rectangular) distribution ranging from 0.0 to 1.0 and compared to $P(\theta)_j$. If X_j is larger than $P(\theta)_j$, then an incorrect response is specified for the item; otherwise, a correct response is specified for the item. Thus, a "subject" with $P(\theta)_j = .90$ gets the item correct 9 in 10 times, and a vector of item responses is developed for each "subject" in each data set. These response vectors are then used to investigate the AIM linking procedures.

Table 2 shows the distribution of ICC parameters for the 80 items for Test 1 (T1) and Test 2 (T2), while Table 3 shows the ICC parameters for the 20 anchor items which are common to both tests.

Subjects from Group 1 were administered only the items in Test 1, and subjects from Group 2 only the items in Test 2. In order to study the effects of sample size, the ICC parameters were estimated on four samples drawn with replacement as follows: 250; 500; 1,000; and 2,000. The ICC parameters were estimated on these four sample sizes for both groups. Anchor ICC parameter values from the four samples administered Test 1 serve as the input values for the anchor item parameters to the second test. This permitted the four sizes of calibration sample (250; 500; 1,000; 2,000) to be varied and tried out with the four samples used to estimate the anchor item ICC parameters.

Table 2. **Means and Standard Deviations of the Generated Item Parameters for Test 1 (T1) and Test 2 (T2)**

Parameter	Test T1	Test T2
$\bar{a}$	1.0564	1.0452
σ_a	0.2793	0.2394
$\bar{b}$	0.0847	−0.0559
σ_b	0.8442	0.8577
$\bar{c}$	0.1878	0.2017
σ_c	0.0542	0.0474

Table 3. **ICC Parameters of the 20 Anchor Items Common to Both Tests**

Number	ICC Parameter		
	a	b	c
1	.8000	−1.5000	.1000
2	.8000	−1.3500	.1000
3	1.0000	−1.2000	.1500
4	1.0000	−1.0500	.1500
5	1.1000	−.9000	.2000
6	1.2000	−.7500	.2000
7	1.2000	−.6000	.2200
8	1.2000	−.4500	.2000
9	1.3000	−.3000	.2000
10	1.4000	−.1500	.2000
11	1.4000	.1500	.2200
12	1.3000	.3000	.2500
13	1.2000	.4500	.2000
14	1.2000	.6000	.2200
15	1.1000	.7500	.2200
16	1.0000	.9000	.2000
17	1.0000	1.0500	.2500
18	.8000	1.2500	.2500
19	.8000	1.3500	.2500
20	.8000	1.5000	.2500
Mean	1.0600	.0000	.2015
S D	.2113	.9549	.0453

III. RESULTS

Table 4 shows the intercorrelations between the known item parameters and the estimated parameters. As past research indicates (Urry, 1976), the correlations all increase with increasing sample size. The correlations in Test 1 for *b* and estimates of *b* start high at .952 and increase to an exceptionally high .992. Correlations for *a* and estimates of *a* begin moderately at .666 and climb to .869, but the correlations of *c* and estimated *c* increase from only .031 to .115. In Test 2, much the same pattern is observed except that the correlation of *c* and estimated *c* increases from .164 to .315 as sample size increases.

Because correlations are insensitive to constant differences as might be found if ICC parameters are overestimated or underestimated by a constant amount, summed absolute deviates of the estimated parameters from the known parameters were computed for each parameter in each sample size. Table 5 presents the summed absolute deviations (or summed errors) for both tests with the four sample sizes. Figure 2 displays this graphically. There is a large drop in summed error when the *a* parameter is estimated on progressively larger samples of subjects up to and including the difference between 1,000 and 500 subjects. Between 1,000 and 2,000 subjects, the difference in summed error is smaller. The relationship between error and sample size for the *b* parameter is more nearly constant. That is, the line on the figure for estimates of *b* is generally straight which means error tends to be reduced in direct proportion to the number of subjects. The almost flat line for the *c* parameter indicates that virtually no reduction of error is occurring

Table 4. **Intercorrelations between Known and Estimated ICC Parameters for Both Groups with Varying Sample Sizes**

Parameter	N	Test 1	Test 2
a	250	.666	.512
	500	.671	.725
	1,000	.831	.813
	2,000	.869	.886
b	250	.952	.929
	500	.964	.962
	1,000	.980	.979
	2,000	.992	.987
c	250	.031	.164
	500	.035	.109
	1,000	−.012	.331
	2,000	.115	.315

Table 5. **Summed Absolute Deviations ($\Sigma|\text{Error}|$) and Average Absolute Deviations ($\overline{|\text{Error}|}$) for the Three ICC Parameters for the Two Tests**

		Test 1		Test 2	
Parameter	N	$\Sigma\|\text{Error}\|$	$\overline{\|\text{Error}\|}$	$\Sigma\|\text{Error}\|$	$\overline{\|\text{Error}\|}$
a	250	30.6450	.3831	30.5290	.3816
a	500	22.8090	.2851	20.6910	.2586
a	1,000	15.7490	.1969	16.8910	.2111
a	2,000	15.5980	.1950	15.1390	.1892
b	250	23.5050	.2938	20.8470	.2606
b	500	19.8600	.2483	16.6070	.2076
b	1,000	17.6890	.2211	13.8050	.1726
b	2,000	12.7350	.1592	11.5130	.1439
c	250	7.7360	.0967	7.2350	.0904
c	500	7.3600	.0920	7.5120	.0939
c	1,000	6.9080	.0864	7.3180	.0915
c	2,000	6.4400	.0805	6.8640	.0858

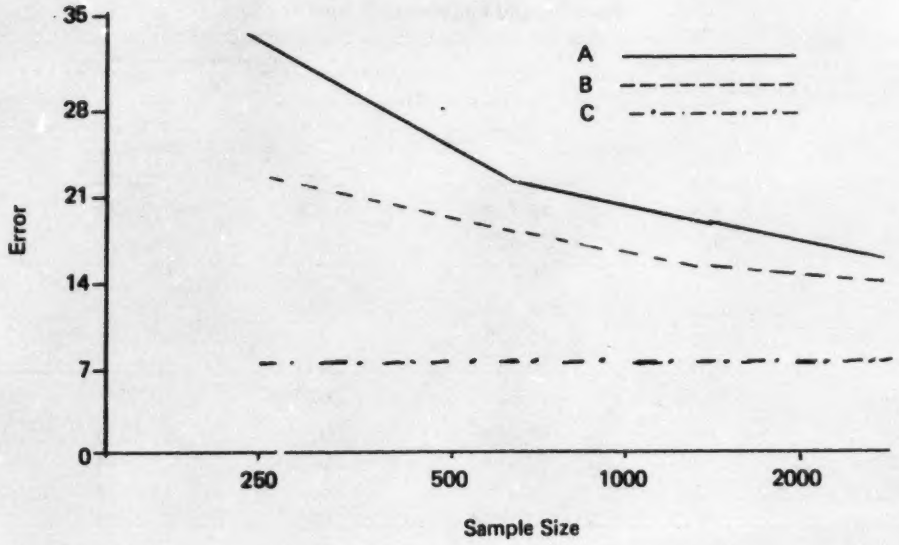

Figure 2. **Error in Estimation of ICC Parameter.**

with increasing sample size for that parameter. The average absolute deviation for the *c* parameter is almost one-third of the entire range of the parameter as the *c* parameter is generally estimated between .00 and .30. However, past research (Ree, 1979) indicates that, even for low ability subjects, the effects of errors in the estimation of the c parameter are small.

Summed deviations of known ICC parameters from the equated value of the ICC parameters were computed for the *a* and *b* parameters for the 16 combinations of calibration sample size and equating sample size. Table 6 shows the summed deviations and the per item deviation for both parameters for the 16 combinations. The equated *a* parameter shows large summed deviations whenever the sample has been limited to 250 subjects whether in the calibration or equating sample. The lowest error rates for the a parameter occur when the anchor item values have been estimated on 2,000 subjects. The effects of the size of the calibration sample are not so clear-cut. When 2,000 subjects are used to estimate the anchor item ICC parameters, the magnitude of the error is approximately the same for all calibration sample sizes except 250. With increasing calibration sample size, the error rate increases by some small amount as indicated by the average (per item) error. This is an unexpected result and an explanation may be found in the relationship between the sets of estimated a parameters. If the estimated *a* parameters were all estimates of the same value and if the test scale were unidimensional, a basic assumption of the theory, then the estimated *a* parameters should be linear transformations of one another and should be correlated 1.0, as correlations are invariant to a linear transformation. This was not found to be the case, and Table 7 shows the intercorrelation of the estimated *a* parameters. Only the correlation between the estimate of *a* calculated on 1,000 subjects and the estimate of *a* calculated on 2,000 subjects approaches this relationship. This lack of linearity may be due to the assumption of normality and to the rescaling used in the calibration procedure, and these may interact in such a way as to produce the anomalous results. Table 8 shows the intercorrelation of estimated *b* parameters. All exceed .900, and the summed deviations also show a steady decrease as sample size increases for the *b* parameter, indicating a virtually linear transformation of

Table 6. **Summed Absolute Deviations ($\Sigma|Error|$) and Average Absolute Deviations ($\overline{|Error|}$) for the a and b Parameters for Various Equating and Calibrating Sample Sizes**

Number of Subjects		Parameter			
		a		b	
Calibration	Equating	$\Sigma\|Error\|$	$\overline{\|Error\|}$	$\Sigma\|Error\|$	$\overline{\|Error\|}$
250	2000	34.2263	.4278	23.3679	.2921
500	2000	15.1282	.1891	21.9342	.2742
1000	2000	15.9871	.1998	16.3660	.2046
2000	2000	16.5958	.2074	13.4579	.1682
250	1000	38.3625	.4795	25.6440	.3205
500	1000	17.6788	.2210	24.3413	.3043
1000	1000	19.5867	.2448	19.1156	.2389
2000	1000	21.0321	.2629	16.8828	2110
250	500	48.6112	.6076	25.4374	.3180
500	500	24.5582	.3070	22.8994	.2862
1000	500	28.8291	.3604	18.1871	.2273
2000	500	31.2094	.3901	15.8328	.1979
250	250	44.3122	.5539	26.2011	.3275
500	250	21.5767	.2697	24.4160	.3052
1000	250	24.4389	.3117	19.4843	.2436
2000	250	27.0242	.3378	17.3255	.2166

Table 7. **Intercorrelations, Means, and Standard Deviation of the Estimated a Parameters[a] for Test 2**

	1	2	3	4
1	1.000			
2	.757	1.000		
3	.690	.860	1.000	
4	.595	.803	.926	1.000
Mean	1.3525	1.2539	1.2348	1.2268
SD	.4843	.3347	.3254	.3061

[a]Variables are for the four sample sizes: 250; 500; 1,000; 2,000.

Table 8. **Intercorrelations, Means, and Standard Deviation of the Estimated b Parameters[a] for Test 2**

	1	2	3	4
1	1.00			
2	.952	1.00		
3	.940	.978	1.00	
4	.935	.969	.986	1.00
Mean	.0563	.0591	.0735	.0559
SD	.8558	.8384	.8700	.8727

[a]Variables are for the four sample sizes: 250; 500; 1,000; 2,000.

estimated *b* parameters from sample to sample. However, with 500 subjects in the equating sample, a similar anomaly is observed which may also be due to normal assumptions and to rescaling.

IV. DISCUSSION

The results of the study present new evidence of the critical interrelationship between item calibration and equating sample sizes and the values of ICC parameters.

Estimating and Equating a

For the 16 combinations of calibration sample sizes and equating sample sizes identified in Table 6, the least deviation of estimated *a* from its known value occurred with an equating sample size of 2,000 and a calibration sample size of 500. As mentioned in the previous section, although the least error between the estimated and known a values was expected with a match of 2,000 equating and 2,000 calibrating sample sizes, the error actually increased very slightly with increasing calibration sample sizes beyond 500. This discrepancy apparently results from a non-linear transformation with sample sizes of 250 and 500 but tends toward linearity with sample sizes of 1,000 and 2,000.

During equating procedures, a sample size $>$ 500 should be developed to ensure an acceptable degree of confidence that the estimation of *a* does not significantly depart from its "true" value. In the same light, estimation of *a* suffers considerably using equating sample sizes of less than 500 such that equating samples of 1,000 or 2,000 are highly desirable to minimize error in estimating *a*.

Estimating and Equating b

Table 6 also shows the linear relationship between error and sample size for the b parameter. The *b* parameter is best estimated with calibration and equating samples of 2,000 each, although a calibration sample size of 1,000 with an equating sample size of 500 can be tolerated without an appreciable increase in error. With all combinations of calibration and equating sample sizes, *b* is estimated quite well.

Estimating and Equating c

The flat line drawn in Figure 2, representing the data from Table 5, shows the estimation of the c parameter to be nearly insensitive to increases in sample size. As sample size increases from 250 to 2,000 subjects, the error decreases but only very slightly. With the c defined as the lower asymptote of the ICC and representing the probability of extremely low ability examinees correctly answering an item, the inability to estimate c with precision could be disturbing. However, it has been pointed out (Lord, 1975) that if a $(\theta - b) < -2$, then the probability of a correct response is c. Therefore, if there are a large number of subjects with ability θ so that $\theta < -(2/a - b)$, c can be accurately estimated. If this requirement is not met, c will be poorly estimated.

A stable and accurate estimate of the a and b parameters requires large numbers of subjects over a broad range of ability. The estimation of c requires large numbers of subjects at very low ability levels. This holds for both equating and calibrating samples; therefore, it is necessary to administer test items, whether to be calibrated or equated, to the largest samples available.

REFERENCES

Lord, F.M. *Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters.* ETS-RB-75-33. Princeton, NJ: Educational Service, 1975.

Lord, F.M. Practical applications of item characteristic curve theory. *Journal of Educational Measurement,* 1977, **14**, 117-138.

Lord, F.M., & Novick, M.R. *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley, 1968.

Ree, M.J. *Adaptive testing at an AFEES.* Paper delivered at the Annual Meeting of the Military Testing Association, San Antonio, Texas, 1977.

Ree, M.J. *Estimating item characteristic curves.* AFHRL-TR-78-68, AD-A064 739. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, November 1978.

Ree, M.J. *The effects of errors in the estimation of item characteristic curve parameters.* Paper presented at the Annual Meeting of the Military Testing Association, San Diego, California, 1979.

Sympson, B., & Ree, M.J. *A validity comparison of adaptive testing in a military technical training environment.* Brooks AFB, TX: Air Force Human Resources Laboratory, Manpower and Pesonnel Research Division, in press.

Urry, V. A five-year quest: Is computerized adaptive testing feasible? *Proceedings of the First Conference on Computerized Adaptive Testing,* Washington D.C.: Government Printing Office, March 1976.